


Measuring Effectiveness: What Will It Take?

As states look to a more active role in teacher evaluation, they face seven major challenges.



Circe Stumbo and Peter McWalters



As the dust settles from the flurry of activity surrounding the education stimulus package and the new programs it created—such as Race to the Top, Investing in Innovation, and the School Improvement Grants—a clear message has taken shape: Federal policy now focuses on teacher “effectiveness” rather than teacher “quality.”

The centerpiece federal law for K–12 education, the Elementary and Secondary Education Act (ESEA), set teacher quality as a major policy priority when it was reauthorized as No Child Left Behind (NCLB) in 2001. *Teacher quality* largely refers to how well teachers know their content as measured by the postsecondary courses they have taken. The shift toward *effectiveness* focuses on how well teachers perform with students. Rather than measuring inputs (such as how many academic degrees the teacher has or how long he or she has been on the job), we should measure the outcomes of a teacher’s work to see how effective the teacher is (the extent to which the educator has met crucial student needs, such as increasing student achievement). This is analogous to the shift from paying attention to student inputs (how many courses a student has taken, or “seat time”) to looking at outcomes (how much the student knows and can do, or performance).

Recent Advances

Although measuring outcomes rather than inputs has been the expressed intention of standards-based reforms for at least two decades, policy changes that make that shift real

have been slow to come to fruition. The year 2010 sped up the pace of reform. The new attention to effectiveness is most obvious in the call for improving teacher evaluation. Although evaluation has traditionally been a local responsibility, federal programs have been calling for states to require evaluation systems that include specific measures of teacher effectiveness, such as student achievement data.

For example, section (D)(2)(ii) of the Race to the Top application (U.S. Department of Education, 2009) asks states to “design and implement rigorous, transparent, and fair evaluation systems for teachers and principals that . . . differentiate effectiveness using multiple rating categories that take into account data on student growth . . . as a significant factor” (p. 34). Although there is no clear definition of “significant,” some of the winning Race to the Top states set the weight of student performance at 50 percent or more of a teacher’s evaluation score.

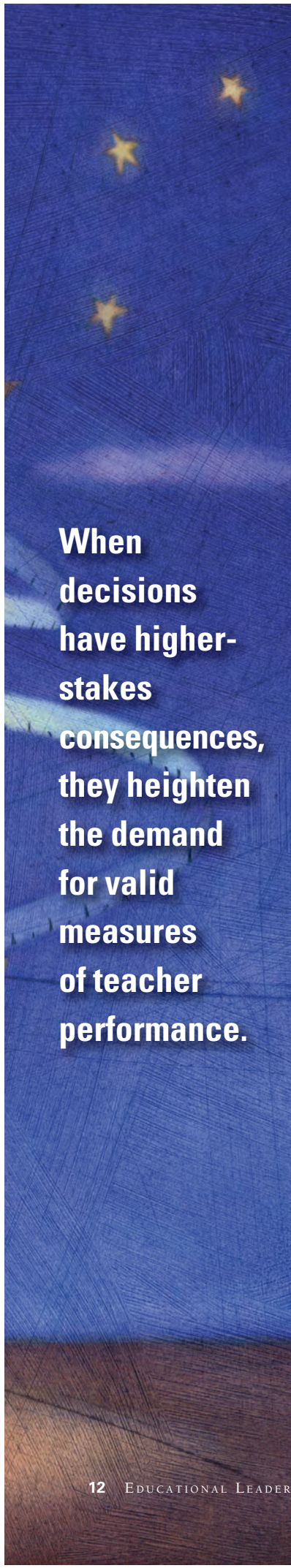
The focus on teacher effectiveness does not end with the stimulus fund programs. The administration’s *A Blueprint for Reform* (U.S. Department of Education, 2010) lays out proposals as Congress engages in its periodic review of ESEA:

We are calling on states and districts to develop and implement systems of teacher and principal evaluation and support, and to identify effective and highly effective teachers and principals on the basis of student growth and other factors. (p. 4)

Through ESEA, the effectiveness agenda could become enduring policy.







**When
decisions
have higher-
stakes
consequences,
they heighten
the demand
for valid
measures
of teacher
performance.**

Seven Challenges Ahead

Proposals for change in the way we evaluate teachers are particularly knotty when teacher evaluation is connected to high-stakes decisions such as tenure, promotion, removal, or compensation. As part of each teacher's regular evaluation, some districts already look at student test score data that demonstrate how much students advanced while working with that teacher (value-added data). However, districts tend to use that information to determine things like a teacher's professional development needs.

When decisions have higher-stakes consequences, such as a teacher potentially losing her or his job, they heighten the demand for measures of teacher performance that are valid, reliable, credible, fair, and legally defensible. As states move to take a more active role in teacher evaluation, they face major challenges in meeting this demand. These challenges raise questions that range from the psychometrics of creating valid and reliable measures of effectiveness to the purpose of public education.

Challenge 1: The Limits of Student Assessment Data

Sophisticated value-added modeling—using student assessment data, adjusted for some student and school characteristics, to determine how much growth in student performance occurred with a particular teacher—is relatively untested as a high-stakes measure, as demonstrated by the controversy that arose when the *Los Angeles Times* released value-added assessment data by teacher (see <http://projects.latimes.com/value-added/>). According to highly regarded testing experts, the evidence supporting the validity and reliability of value-added modeling results is weak enough that such results should not yet be used as the major measure of teacher effectiveness (Baker et al., 2010).

Similarly, testing experts such as W. James Popham and members of the Commission on Instructionally Supportive Assessment argue that the types of standardized exams used in most value-added assessment systems are not “instructionally sensitive.” Popham (2007) defines “instructional sensitivity” as “the degree to which students’ performances on a test accu-

ately reflect the quality of instruction specifically provided to promote students’ mastery of what is being assessed.”

Unfortunately, too many standardized exams do *not* demonstrate whether a teacher’s instruction had an effect on the student’s performance. With this kind of assertion waiting in the wings, educators are likely to challenge the use of summative statewide standardized exams in high-stakes evaluations.

Challenge 2: Many Untested Subjects

The most obvious problem associated with attributing individual teachers’ performance to individual students’ test scores is connecting test scores to teachers who teach untested subjects. Every state administers English language arts and mathematics tests in grades 3–8 as well as once in high school. Thus, preK–2 and three years in high school are mostly untested. Further, although many states administer tests in science and social studies, they are not administered at every grade level and may not provide the right kind of information for teacher evaluation in those subject areas.

We can see how these difficulties play out if we look at a student in a traditional junior high school. Last year, Jacob, the son of one of the authors, was a 7th grader. His favorite class was Global Studies. In 7th grade, students are required to take two full-year courses in literacy and language arts and just one Global Studies course for one trimester. The Iowa Tests of Basic Skills (ITBS) cover social studies. However, as the ITBS website states, “The content of the [social studies] questions is taken from the areas of geography, history, government, economics, sociology, and the other social sciences” (Iowa Testing Program, n.d.). Is it reasonable to believe that Jacob’s Global Studies teacher, with just 60 days with each child, stands a chance at preparing students to succeed on such a general exam? To what extent are Jacob’s scores on the ITBS an accurate measure of his teacher’s performance?

It is possible to find alternative measures of student performance that can be compared across classrooms beyond statewide, multiple-choice standardized exams, such as the National Writing Project’s rubrics and juried competitions to judge senior year capstone

projects for graduation. Some states (such as Vermont, Maine, Rhode Island, and Nebraska) and high-performing countries (such as Finland) are engaged in this work, but it is both complicated and expensive.

If we are using such student performances for high-stakes decision making, we will need to make sure that the determinations are valid (that we measure what we mean to measure)

Brown-Sims, & Hess, 2007; Danielson & McGreal, 2000; Little, Goe, & Bell, 2009).

Lack of evaluator training is a threat to the reliability of the evaluation and objectivity of the results. An untrained observer may introduce bias into observations; the observer's expectations of a teacher may influence the observation to a greater degree than the actual teacher behaviors displayed (Mujis, 2006). Researchers at the National Comprehensive Center for Teacher Quality (Little et al., 2009) found that a "reliable classroom observation protocol may be wildly inaccurate or inconsistent in the hands of an untrained evaluator" (p. 21).

However, districts rarely require evaluators to be trained. Only 8 percent of the districts in a study of midwestern districts had written documentation detailing requirements for training evaluators (Brandt et al., 2007; Loup, Garland, Ellett, & Rugutt, 1996). In many of the 12 districts examined in the *Widget Effect* (Weisburg, Sexton, Mulhern, & Keeling, 2009), evaluation training was a one-time endeavor provided either when an administrator was new to the position or when the district implemented a revised system of teacher evaluation. Inter- and intra-rater reliability is also increasingly needed as evaluations inform high-stakes decision making, but this has not been developed yet (Joint Committee on Standards for Educational Evaluation, 1994; Mathers, Oliva, & Laine, 2008; Mujis, 2006).

Recognizing this challenge, Iowa, Minnesota, and North Carolina are providing training for evaluators, but few other states have taken up a similar mantle. We predict that evaluator training will become a priority for many states.

Challenge 4: Individual Versus Team-Based Accountability

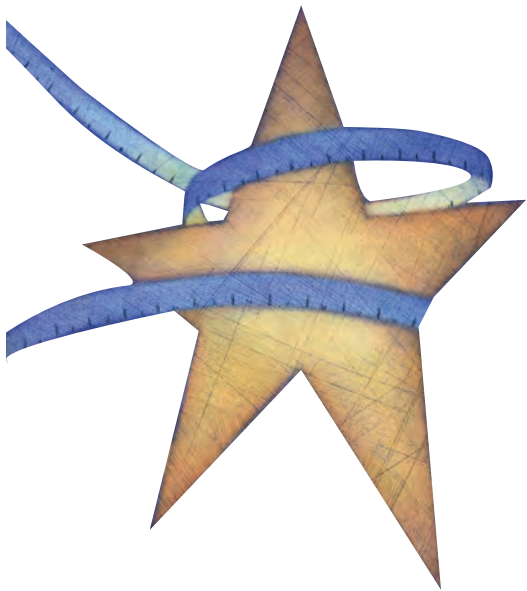
It's difficult to attribute student perfor-

mance to a specific teacher in secondary school or in virtual programs because students in these environments have multiple teachers daily. But even in elementary school, which traditionally assigns students to one teacher only, students who need additional learning supports might work with an adult in addition to the teacher of record on basic skills in English language arts or mathematics.

Let's look again at Jacob's junior high school, which is organized by trimesters. Jacob had 11 different courses last year, meaning 11 different teachers. All 7th graders in his school take two courses in English language arts. Which of his two English teachers can claim success with Jacob on the basis of his language arts scores on the Iowa Tests of Basic Skills? Moreover, can his Global Studies teacher lay claim to success in any of those areas—or is social studies the only test that matters to her? The Iowa Testing Program states, "The questions on this [social studies] test measure objectives of the social studies curriculum that are not measured elsewhere in the ITBS tests" (Iowa Testing Program, n.d.). This implies that her instruction may have an effect on Jacob's scores on other tests, but it's unclear how to discern that effect.

Trying to attribute student performance to a specific teacher also runs counter to the collaborative way we think about teaching today. Teachers who collectively engage in participatory decision making, lesson design, data analysis, and analysis of student work are better able to deliver rigorous and relevant learning for all students and personalize learning for individual students. The new core teaching standards reflect this understanding, calling for teachers to participate actively as team members in decision-making processes.

Most teacher evaluation systems have been designed to assess individuals, but the collaborative culture envisioned by the new core teaching standards (and



and reliable (that the measure will yield the same results on repeated trials). For the arts, physical education, and other untested subjects, the development of such measures of student performance has yet to be completed.

The federal Race to the Top program acknowledged this challenge. In both the state and assessment consortium grants, it provided funding that could be used to improve the type and quality of assessments. Still, this is a massive undertaking.

Challenge 3: Quality of Evaluators

A scan of the literature on teacher evaluation demonstrates that teachers do not routinely and consistently receive quality evaluations. Several studies examine deficiencies in administrators' ability to conduct quality evaluations (see Brandt, Mathers, Oliva,



by the administration's reauthorization blueprint, for that matter) will require us to explore a next-generation, team-based approach to performance review.

Challenge 5: What Else Matters?

Up to now, under NCLB, teachers have defined teacher quality as knowing their subject matter. The Council of Chief State School Officers (CCSSO) has recently revised its model core teaching standards, which go beyond possessing content knowledge to incorporate knowledge of *how to teach* one's subject matter (for example, how to identify students' common misunderstandings and help students move beyond them) and "how to connect concepts and use differing perspectives to engage learners in critical/creative thinking and collaborative problem solving related to authentic local and global issues" (CCSSO, 2010, p. 15). Presumably, students whose teachers have this set of skills would perform well on exams.

But even in the most stringent state policy propositions, one-half of a teacher's evaluation is based on criteria other than student performance. How a teacher helps students to become motivated to learn, persist in their work, strive to be lifelong learners, express themselves artistically, behave civilly, and not bully others—these factors matter to parents, students, and communities. The Obama administration captures these sentiments in its call for a more holistic understanding of education. States have asserted their interest in citizenship education, not just college and career readiness. And teachers point to an obligation to support a culture of learning in their school communities as well as to develop their profession. As the focus on teacher evaluation rises to the state and federal levels, we will need to articulate the full range of teacher practices and student outcomes that we want from our education system—and determine how we can measure them.

Challenge 6: Working Conditions

When developing an approach to teacher evaluation and its high-stakes consequences, states will need to consider the systems in which teachers work. Have teacher evaluation systems taken into account circumstances beyond teachers' control? These range from having access to appropriate resources (such as a heated classroom) or equipment that enhances learning (such as computers); to access to professional communities of support (such as other teachers with whom to collaborate, behavior specialists, and other resource staff); to the alignment of education programs among the school, district, and state.

And what about other conditions over which teachers have little control, such as student readiness? Are students hungry or suffering? Is the school climate conducive to student learning and teacher collaboration? Research suggests that improved working conditions significantly influence a school's ability to reach achievement goals (for a full summary, see Emerick, Hirsch, & Berry, 2005), yet we have few strong models that account for working conditions in evaluating teacher effectiveness.



Challenge 7: Engaging All Stakeholders

Teacher evaluation has primarily been a local responsibility, but federal programs such as Race to the Top signal a shift toward using evaluation to meet state and federal goals. Aligning these multiple levels of authority to support the dual purposes of evaluation—professional growth and accountability—will require adjusting the purpose, design, and mechanisms of evaluation systems. It will also require a culture of shared responsibility and mission as more players claim a stake in the outcomes of teacher evaluation and take a more active role in designing evaluation systems.

No matter how we proceed, we need to engage all stakeholders in the discussion. Stakeholders include professional standards boards, boards of examiners, professional organizations, membership associations, unions, boards of regents, teacher educators, professional developers, local school boards, and teachers and administrators. Only then can we clarify feasibility, mobilize interest, anticipate and prevent barriers, and ensure high-fidelity implementation.

What Next?

We raise these challenges not to sound an alarm, but to suggest an agenda for cooperative research, design, development, and assessment of state policy and local practices. When grounded in agreed-on standards for teaching—such as the recently revised Interstate Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards—and developed in ways that overcome the challenges cited, evaluation can be an effective lever for state and local policy.

Having just completed the initial stage of revising the standards, CCSSO is now turning its attention to crafting a developmental continuum across a teacher's career, pegged to the standards, as well as rubrics for evaluating



progress. An additional goal is to work to validate the standards in practice. We expect these efforts, as well as the efforts of states already using the draft standards in their teacher evaluation pilots, will lead to a rigorous foundation for a system of teacher performance.

We further expect the assessment community to work vigorously to design multiple measures of student performance in both tested and untested subjects and to develop value-added assessment systems that have greater reliability. Two assessment consortia received funding through Race to the Top to engage specifically in this work. In addition, groups such as the Teacher Performance Assessment Consortium—a partnership among the American Association of Colleges of Teacher Education, the CCSSO, and Stanford University—are developing improved teacher performance assessments and portfolio systems. Projects such as the Measures of Effective Teaching, developed by the Bill and Melinda Gates Foundation, are evaluating video samples of practicing teachers against validated rubrics, student surveys, and student performance.

Finally, we will need to learn from experts in the business community, who have long been working on team-based accountability systems, how to shift the model from the individual as the sole unit of authority and responsibility to next-generation systems that recognize the importance of professional collaboration, transparent practice, reflective and collective inquiry, and joint accountability. ■

References

- Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R. L., et al. (2010). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper No. 278). Washington, DC: Economic Policy Institute.
- Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). *Examining district guidance to schools on teacher evaluation policies in the Midwest Region* (Issues & Answers Report, REL 2007–No. 030). Washington, DC: U.S. Department of Education, Regional Educational Laboratory Midwest. Retrieved October 16, 2008, from http://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_2007030.pdf
- Council of Chief State School Officers. (2010). *Interstate Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards* (Draft for public comment). Washington, DC: Author.
- Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation: To enhance professional practice*. Alexandria, VA: ASCD.
- Emerick, S., Hirsch, E., & Berry, B. (2005). Teacher working conditions as catalysts for student learning. *InfoBrief*, 43. Alexandria, VA: ASCD.
- Iowa Testing Program. (n.d.). *Description of Iowa Tests of Basic Skills Tests, Levels 5–8 (Primary Grades)*. Retrieved from www.education.uiowa.edu/itp/itbs/itbs_about_5-8.aspx
- Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.
- Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved April 25, 2010, from www.tqsource.org/publications/practicalGuide.pdf
- Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest districts. *Journal of Personnel Evaluation in Education*, 10(3), 203–226.
- Mathers, C., Oliva, M., & Laine, S. (2008). *Improving instruction through effective teacher evaluation*. Washington, DC: National Comprehensive Center for Teacher Quality.
- Mujis, D. (2006). Measuring teacher effectiveness. *Educational Research and Evaluation*, 12(1), 53–74.
- Popham, W. J. (2007, November). *Instructional sensitivity: Looming challenge for measurement maven*. Presentation at Third Biennial CASMA ACT Conference, Iowa City, Iowa. Retrieved from www.education.uiowa.edu/casma/documents/INSTRUCTIONALSENSITIVITY-Jim.pdf
- U.S. Department of Education. (2009). *Race to the Top application for initial funding*. Washington, DC: Author.
- U.S. Department of Education. (2010). *A blueprint for reform: The reauthorization of the elementary and secondary education act*. Retrieved from www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect*. Brooklyn, NY: New Teacher Project.

Circe Stumbo is president of West Wind Education Policy; circe@westwinded.com.

Peter McWalters is interim strategic initiative director of Education Workforce at the Council of Chief State School Officers and former Commissioner of Education for Rhode Island; peterm@ccsso.org. They are two of the coauthors of *State Policy Implications of the Model Core Teaching Standards and Transforming Teaching and Leading: A Vision for a High-Quality Educator Workforce System*, both of which are available at www.ccsso.org.

The new
attention to
effectiveness
is most
obvious in
the call for
improving
teacher
evaluation.

Copyright of Educational Leadership is the property of Association for Supervision & Curriculum Development and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.